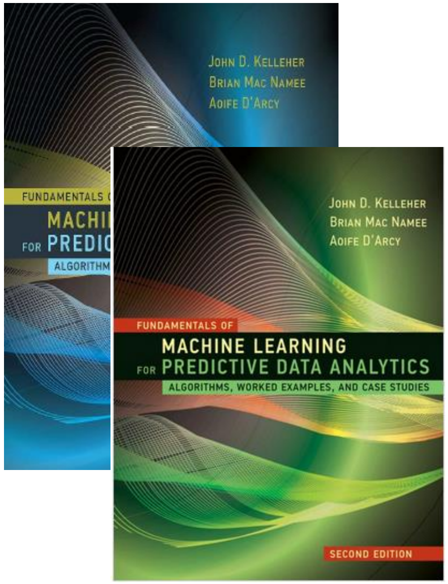
P170M109 Computational Intelligence and Decision Making

Input and Output analysis (part 1)

2022

Fundamentals of Machine Learning for Predictive Data Analytics

...
2. Data to Insights to Decisions
3. Data exploration
....



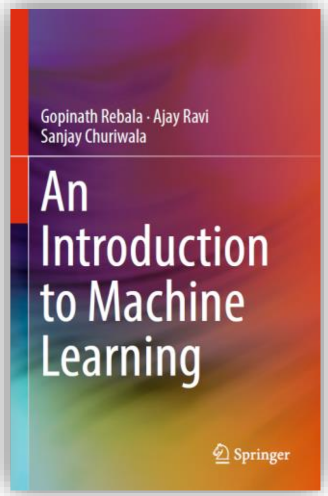
First and Second edition (2015, 2020)

https://mitpress.mit.edu/books/fundamentals-machine-learning-predictive-data-analytics

https://mitpress.mit.edu/books/fundamentals-machine-learning-predictive-data-analytics-second-edition

Literature

Rebala, Gopinath, Ajay Ravi, and Sanjay Churiwala. *An Introduction to Machine Learning*. Springer, 2019.



…
Chapter 6.2 K-Nearest Neighbor (KNN)
Chapter 7, Random Forests

…

https://link.springer.com/book/10.1007%2F978-3-030-15729-6
Use KTU VPN or perform search through https://vb.ktu.edu (uses SSO login and proxy to access full text document)

Predictive Data Analytics

Predictive data analytics is the art of building and using models that make predictions based on patterns extracted from historical data.

Applications:

- Price prediction
- Dosage prediction
- Risk assessment
- Diagnosis
- Document classification
- …

What algorithms are usually applied?

Machine Learning

Machine learning

** an automated process that extracts patterns from data.*
** the science of getting computers to act without being <u>explicitly programmed</u>.*

Supervised machine learning

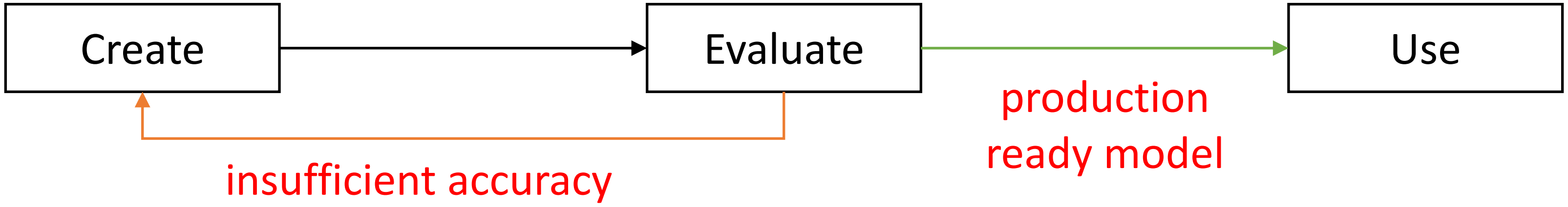
automatically learn a model of the relationship between a set of descriptive features and a target feature based on a <u>set of historical examples</u>.

Supervised Learning

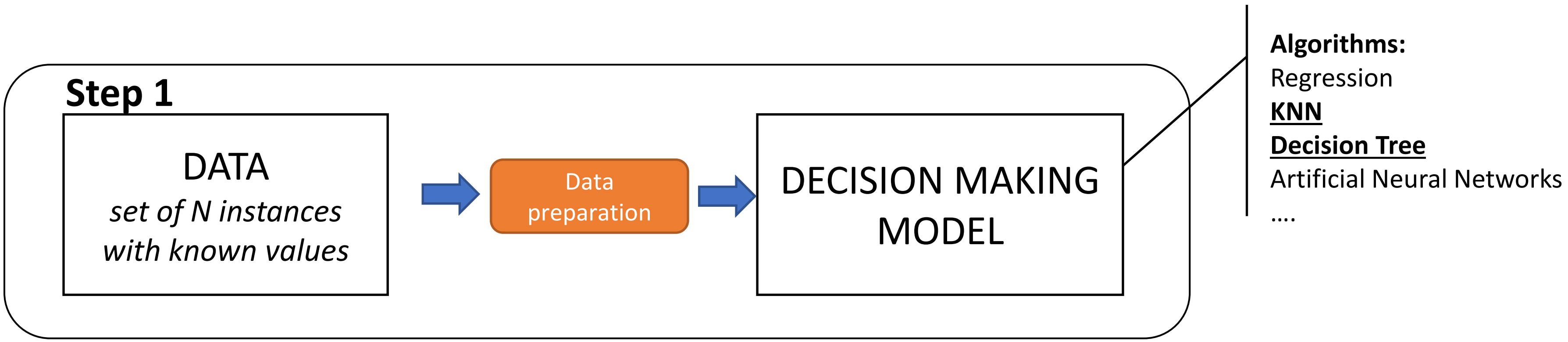
Step 1: designing algorithm based on available data

Step 2: evaluating and improving algorithm based on available data

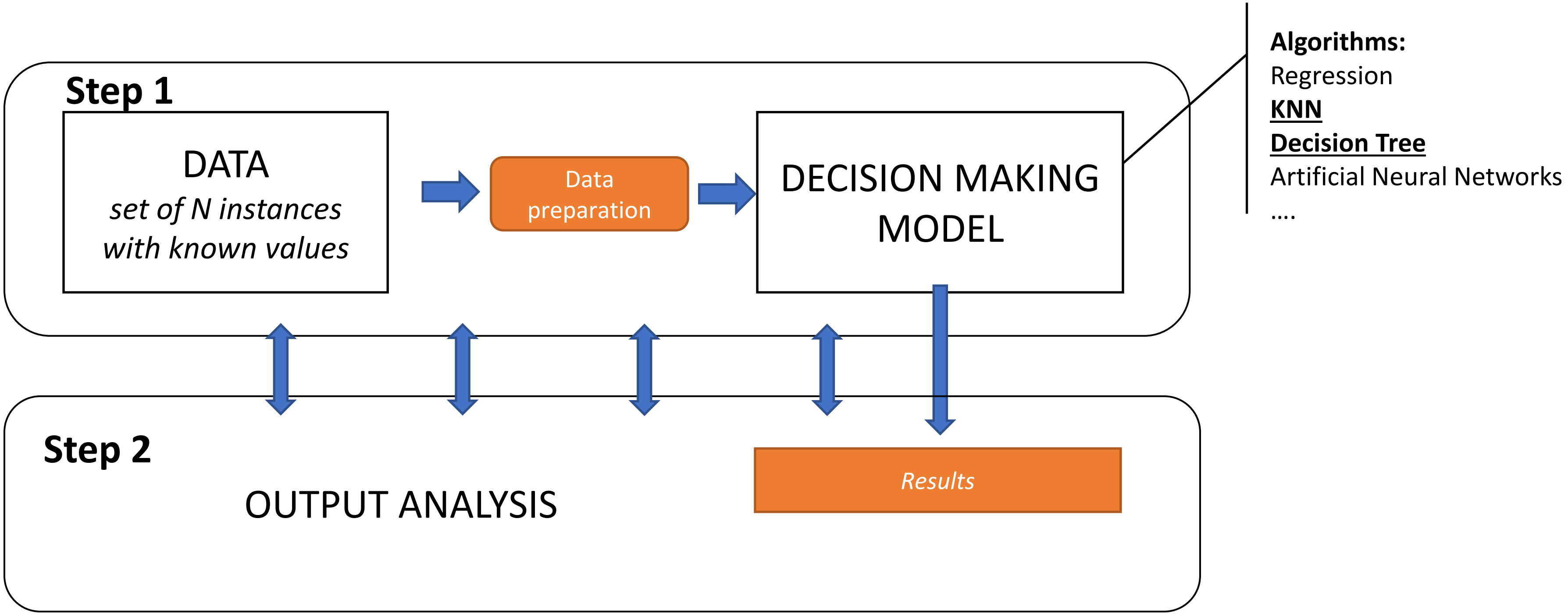
Step 3: using model for the initial problem



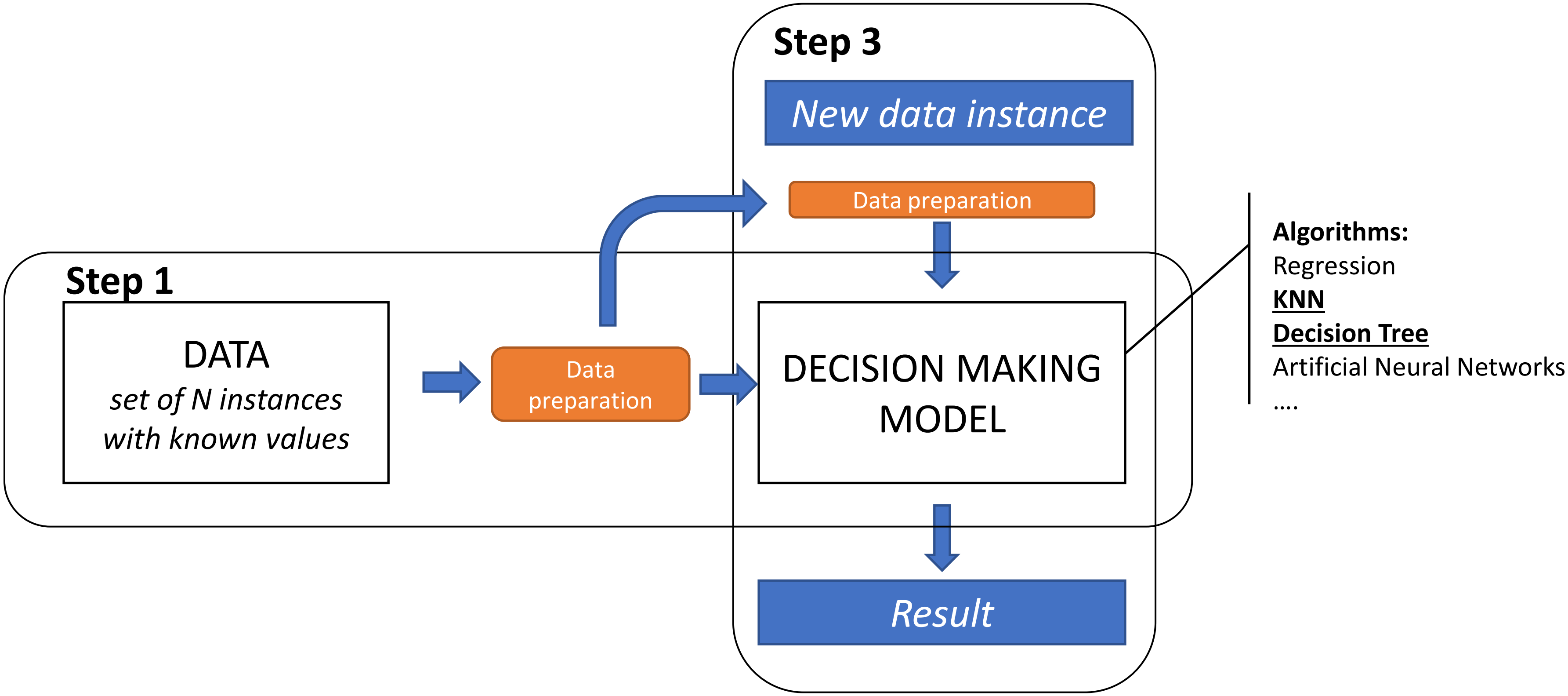
Supervised Learning



Supervised Learning



Supervised Learning



Converting Business Problems into Analytics Solution

1. What is the business problem?
2. What are the goals that the business wants to achieve?
3. How does the business currently work?
4. How a predictive analytics model can help to address the business problem?

Exercise 1

Mobile network company has a business problem:

the increased number of costumers who left the company.

1. How to identify the customers considering switching to a different network provider?
2. Please **provide a list of features** (domain concepts) that could be used for this problem.

3 minutes

Features

- **Raw features**
  come directly from raw data sources
- **Derived features**
  constructed from data in one or more raw data sources
  - **Aggregates** (count, sum, average, minimum, maximum)
  - **Flags** (binary features that indicate presence or absence of some characteristic)
  - **Ratios** (continuous features that capture the relationship between two or more raw data values)
  - **Mappings** (convert continuous features into categorical features)

Why is Data Analysis Important?

let's say, we already have finite dataset

1. To understand characteristics of the data

2. To evaluate data quality:
   - missing values
   - outliers
   - inappropriate level for a feature

Data example

Data for flat price prediction model

Descriptive features						Target feature
Id	**LotArea**	**Surname**	**OverallQual**	**RoofStyle**	**CentralAir**	**Price**
1	8450	Johanson	7	Gable	Yes	150 000
2	9600	Veenker	6	Gable	Yes	150 000
3	11250	Smith	7	Gable	Yes	80 000
4	9550	Crawford	7	Mansard	No	150 000
5	14260	NA	8	Gable	No	220 000
6	14115	Mitchel	5	Gable	Yes	85 000
7	10084	Somerst	8	Gambrel	Yes	200 000
8	NA	Sawyer	7	Gable	Yes	180 000
9	6120	Meadow	7	Gable	Yes	150 000
10	7420	BrkSide	NA	Gable	Yes	50 000
11	11200	Sawyer	5	Hip	Yes	75 000

- *Are all variables important?*
- *Are all variables ready to use?*

Types of Data

- **Numeric**: True numeric values that allow arithmetic operations;
- **Interval**: Values that allow ordering and subtraction, but do not allow other arithmetic operations;
- **Ordinal**: Values that allow ordering but do not permit arithmetic;
- **Categorical**: Values that cannot be ordered and allow no arithmetic;
- **Binary**: A set of just two values;
- **Textual**: Free-form, usually short, text data.

Types of Data

- **Numeric**: True numeric values that allow arithmetic operations;
- **Interval**: Values that allow ordering and subtraction, but do not allow other arithmetic operations;

Continuous

- **Ordinal**: Values that allow ordering but do not permit arithmetic;
- **Categorical**: Values that cannot be ordered and allow no arithmetic;
- **Binary**: A set of just two values;
- **Textual**: Free-form, usually short, text data.

Categorical

Id	LotArea	Surname	OverallQual	DateEvaluated	RoofStyle	CentralAir	Electrical
1	8450	Johanson	7	2016-02-03	Gable	Yes	SBrkr
2	9600	Veenker	6	2016-05-16	Gable	Yes	SBrkr
3	11250	Smith	7	2016-03-12	Gable	Yes	SBrkr
4	9550	Crawford	7	2016-09-25	Mansard	No	SBrkr
5	14260	NA	8	2016-11-13	Gable	No	SBrkr
6	14115	Mitchel	5	2016-10-02	Gable	Yes	SBrkr
7	10084	Somerst	8	2016-02-02	Gambrel	Yes	SBrkr
8	NA	Sawyer	7	2016-07-15	Gable	Yes	SBrkr
9	6120	Meadow	7	2016-02-03	Gable	Yes	FuseF
10	7420	BrkSide	NA	2017-01-01	Gable	Yes	SBrkr
11	11200	Sawyer	5	2017-02-18	Hip	Yes	SBrkr

LotArea: Lot size in square feet;
Surname: Owner’s surname
OverallQual: Rates the overall material and finish of the house (10-very excellent, 1-very poor)

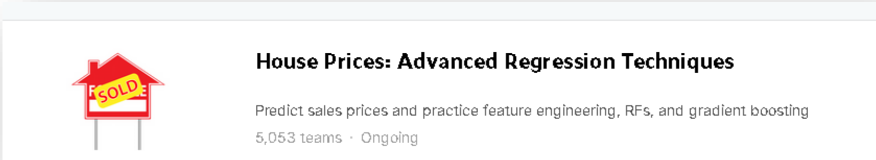
DateEvaluated: The date the quality was evaluated
RoofStyle: Type of roof
CentralAir: Central air conditioning
Electrical: Electrical system

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

RoofStyle: Type of roof (Flat; Gable; Gambrel; Hip; Mansard; Shed)

Electrical: Electrical system (SBrkr *(Standard Circuit Breakers & Romex);* FuseA *(Fuse Box over 60 AMP and all Romex wiring (Average)*); FuseF (*60 AMP Fuse Box and mostly Romex wiring (Fair)*); *FuseP (60 AMP Fuse Box and mostly knob & tube wiring (poor*)); Mix *(Mixed)*)

OverallQual: Rates the overall material and finish of the house (10 (*Very Excellent*), 9 (*Excellent*), 8 (*Very Good*), 7 (*Good*), 6 (*Above Average*), 5 (*Average*), 4 (*Below Average*), 3 (*Fair*), 2 (*Poor*), 1 (*Very Poor*)).



Ordinal	Numerical	Textual	Ordinal	Interval	Categorical	Binary	Categorical
Id	**LotArea**	**Surname**	**OverallQual**	**DateEvaluated**	**RoofStyle**	**CentralAir**	**Electrical**
1	8450	Johanson	7	2016-02-03	Gable	Yes	SBrkr
2	9600	Veenker	6	2016-05-16	Gable	Yes	FuseP
3	11250	Smith	7	2016-03-12	Gable	Yes	SBrkr
4	9550	Crawford	7	2016-09-25	Mansard	No	FuseA
5	14260	Perry	8	2016-11-13	Gable	No	FuseP
6	14115	Mitchel	5	2016-10-02	Gable	Yes	SBrkr
7	10084	Somerst	8	2016-02-02	Gambrel	Yes	FuseA
8	10382	Sawyer	7	2016-07-15	Gable	Yes	SBrkr
9	6120	Meadow	7	2016-02-03	Gable	Yes	FuseF
10	7420	BrkSide	5	2017-01-01	Gable	Yes	SBrkr
11	11200	Sawyer	5	2017-02-18	Hip	Yes	SBrkr

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Data Quality Report

Characteristics of each feature using standard statistical measures of:

- Central tendency:
  - mean
  - median
  - mode
- Variation:
  - standard deviation
  - (visualization) bar plots
  - (visualization) histograms
  - (visualization) box plots

Data Quality Report

Continuous features

Feature	Count	% Miss	Card.	Min	Q1	Mean	Median	Q3	Max	Std. Dev.
---	---	---	---	---	---	---	---	---	---	---
---	---	---	---	---	---	---	---	---	---	---

Categorical features

Feature	Count	% Miss	Card.	Mode	Mode Freq	Mode %	2^{nd} Mode	2^{nd} Mode Freq	2^{nd} Mode %
---	---	---	---	---	---	---	---	---	---
---	---	---	---	---	---	---	---	---	---

Mean (or average)

Given dataset $\{x_1, x_2, \ldots, x_n\}$

$$\boxed{\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

Example:

$\{15; 5; 10; 10;\ 20\}$

$$\mu = \frac{(15 + 5 + 10 + 10 + 20)}{5} = 12$$

Median

The median of a data set is the number that is the middle value of the set.

Given sorted dataset $\{x_1, x_2, \dots, x_n\}, x_i \leq x_{i+1}$

$$median = \begin{cases} x_{\left[\frac{n}{2}\right]+1}, if\ n\ is\ odd \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, if\ n\ is\ even \end{cases}$$

Example 1:

$\{5; 10; 10;\ 15; 30\}$

$median = 10$

Example 2:

$\{5; 10;\ 15; 30\}$

$median = 12.5$

Mode

The mode of a data set is the number that occurs most frequently in the set.

Example 1:

$\{15; 5; 10; 10;\ 20\}$

x_i	No. occurs
5	1
10	**2**
15	1
20	1

Example 2 (bimodal set):

$\{15; 5; 10; 10;\ 20; 6; 20\}$

x_i	No. occurs
5	1
6	1
10	**2**
15	1
20	**2**

Example 3 (no mode):

$\{15; 5; 10;\ 20; 6\}$

x_i	No. occurs
5	1
6	1
10	1
15	1
20	1

Standard Deviation

Given dataset $\{x_1, x_2, \dots, x_n\}$

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}}$$

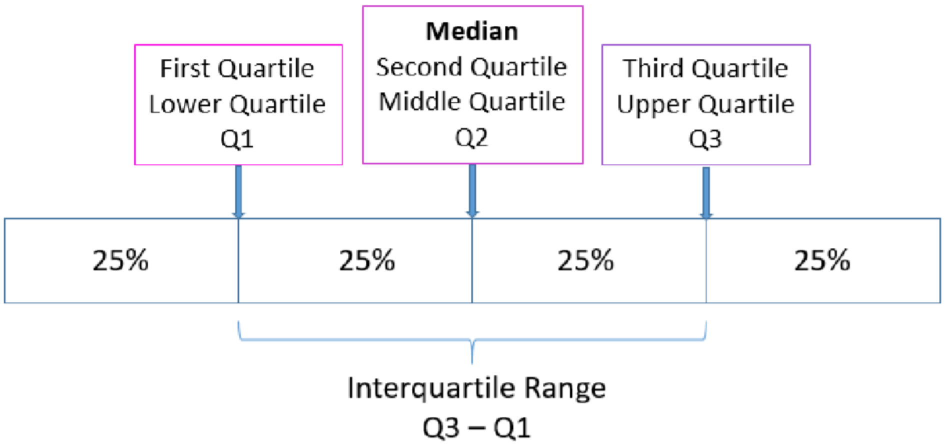
Example:

$\{15; 5; 10; 10;\ 20\}$

$$\mu = \sqrt{\frac{(15-12)^2 + (5-12)^2 + (10-12)^2 + (10-12)^2 + (20-12)^2}{5-1}} \approx 5{,}701$$

Quartile

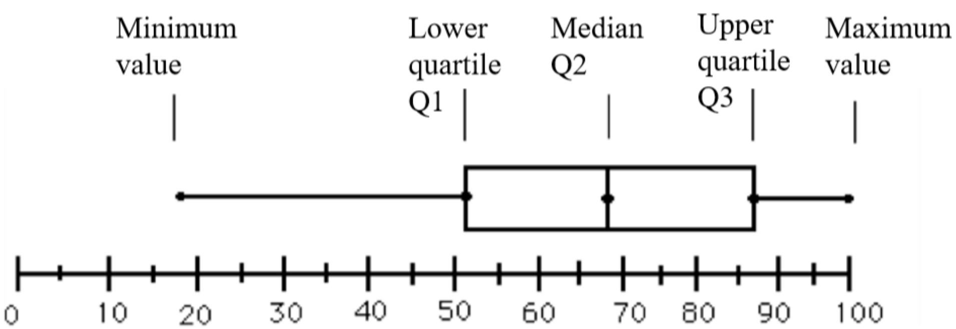
Quartiles are three points that divide sorted data set into four equal groups (by count of numbers), each representing a fourth of the distributed sampled population.

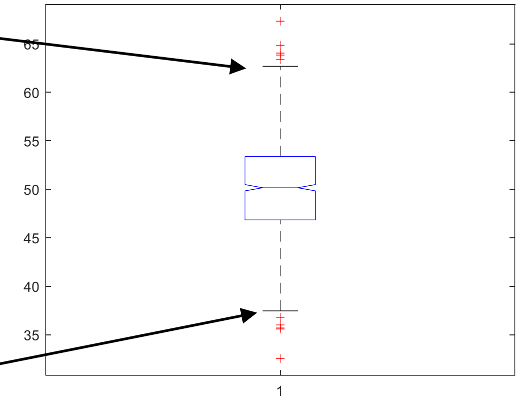


Boxplot

The five-point summary consists of the lower and upper quartiles, the median, the maximum and the minimum values of the data set.

```
Q3 + 1.5 *(Q3-Q1)
```

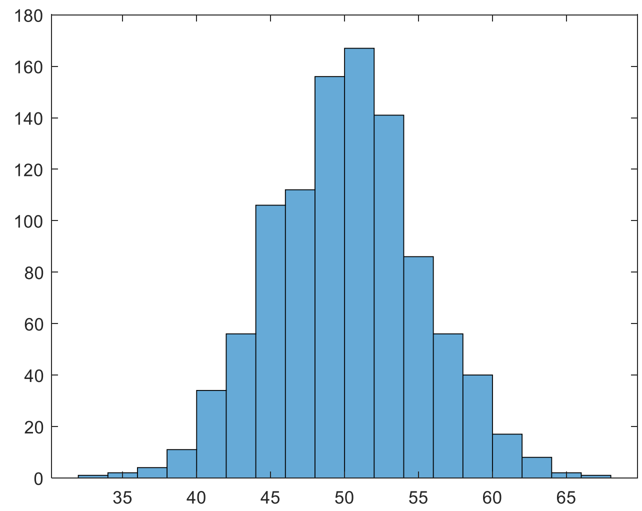





```
Q1 - 1.5 *(Q3-Q1)
```

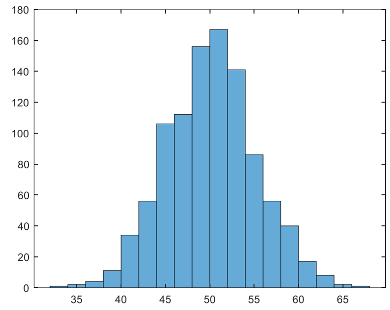
Histogram

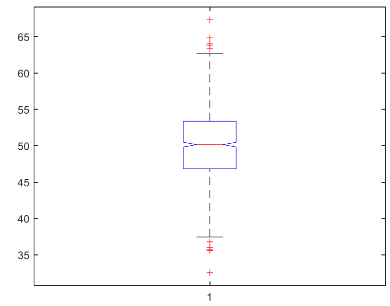
Histograms are a type of bar plot for numeric data that group the data into bins



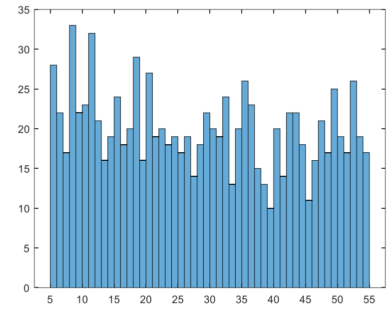
Distributions

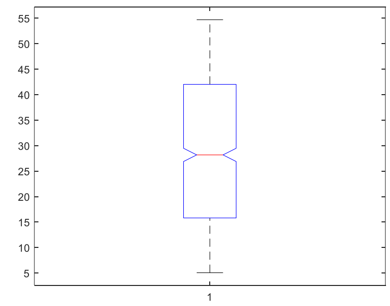
Normal



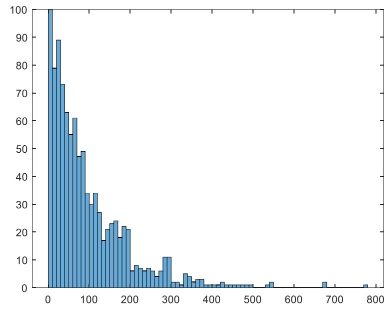


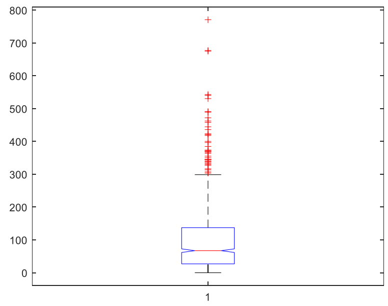
Uniform





Exponential





Data Quality Issues

- missing values
- irregular **cardinality*** problems
- outliers

*Cardinality shows the number of distinct values present for a feature

Data Quality Issues

Missing values

- errors in data integration or in data generation;
- legitimate reasons;
- manual entry;

Remove feature from model if proportion of NA is >60%

Use **imputation**:

- for continuous features replace with mean or median;
- for categorical features replace with mode.

Data Quality Issues

Irregular cardinality problems

Cardinality shows the number of distinct values present for a feature

- cardinality equal to 1
- cardinality of categorical feature is too high
- categorical feature incorrectly labeled as continuous

Data Quality Issues

Outliers

values that lie far away from the central tendency of a feature.

Invalid (noise of data, appeared by mistake) and **valid** (correct values, very different from the rest of the set).

Use **clamp** transformation (clamps all values above an upper threshold and below a lower threshold to these threshold values)

$$x_i = \begin{cases} lower, if\ x_i < lower \\ upper, if\ x_i > upper \\ x_i, otherwise \end{cases}$$

Data Standardization and Normalization

Given dataset $X = \{x_1, x_2, \dots, x_n\}$

Normalization

rescales values into range [0; 1]

$$R = \{r_1, r_2, \dots, r_n\}$$

$$r_i = \frac{x_i - \min(X)}{\max(X) - \min(X)}$$

Standardization (mean normalization)

rescales values to have mean 0 and stdev 1

$$S = \{s_1, s_2, \dots, s_n\}$$

$$s_i = \frac{x_i - \bar{x}}{\sigma}$$

Example:

$\{15; 5; 10; 10;\ 20\}$

Normalizes dataset:

$\{0.667; 0; 0.333; 0.333; 1\}$

Example:

$\{15; 5; 10; 10;\ 20\}$

Standardized dataset:

$\{0.526; -1.228; -0.351; -0.351; 1.403\}$

Example No. 1 Input data analysis

Code example:

- inputDataAnalysis.ipynb
- mixedDataExample.tsv

To do:

- Make analysis of the rest of the variables
- Perform data limitation / model limitation report.
- Fill Data quality reports

	LotArea	OverallQual	YearBuilt	RoofStyle	CentralAir	SalePrice
0	8450	7	2003	Gable	Y	208500
1	9600	6	1976	Gable	Y	181500
2	11250	7	2001	Gable	Y	223500
3	9550	7	1915	Gable	Y	140000
4	14260	8	2000	Gable	Y	250000
...	...	...	...	...	...	...
1455	7917	6	1999	Gable	Y	175000
1456	13175	6	1978	Gable	Y	210000
1457	9042	7	1941	Gable	Y	266500
1458	9717	5	1950	Hip	Y	142125
1459	9937	5	1965	Gable	Y	147500

1460 rows × 6 columns